



LEAD SCORE CASE STUDY

**Presented By,
Navya Mahesh,
Sudhina**

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The goal is to build a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.



BUSINESS OBJECTIVE

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.



ANALYSIS APPROACH

- Since this is a classification problem with identifying if a potential lead is "Hot" lead or not, we employed "Logistic Regression" to tackle this problem.
- First, we understand the data set and apply data pre-processing techniques like dropping insignificant columns, standardizing data, handling missing, outlier and duplicate data.
- We built a models with RFE o/p of 15 variables respectively.
- We compared the evaluation metrics of the 3 models and selected the best one based on the 80% cutoff set by the CEO and also based on the simplicity of the model.



DATA PREPARATION

The following data preparation method was applied so that the data is made dependable to provide significant business value by improving Decision Making Capabilities easier.

Remove columns with only one unique values

'Search','Magazine','NewspaperArticle','XEducationForums','Newspaper','Digital Advertisement','Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content','I agree to pay the amount through cheque', 'A free copy of Mastering The Interview','Country'

Remove rows with high missing values

'Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Index','Asymmetrique Profile Score', 'How did you hear about X Education, Lead Profile

Imputing Null Values with Median

Total Visits and Page Views Per visit are continuous variables with outliers having too many null values Hence they were imputed with median

Imputing Null Values with Mode

Country and City are the categorical columns with null values In country majority of the rows belong to country India and in City majority of rows have information regarding Mumbai Hence they are imputed with India and Mumbai respectively.

DATA PREPARATION

**Handling
Select value in
some column**

As we can observe that there are select values for many column. This is because customer did not select any option from the list, hence it shows select. Select values are as good as NULL. Hence they are converted to NULL.

**Outlier
Treatment**

Total Visits and Page Views Per Visit had some outliers which is being treated by Cap method as the main intention was not to lose any data by dropping values

**Binary
Encoding**

*Converting some binary variables (Yes/No) to 1/0
'Do Not Email', 'Do Not Call'*

**Dummy
Encoding**

For following categorical variables with multiple levels dummy variables were created. 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'Lead Quality', 'City', 'Last Notable Activity'

**Test
Train
Split**

The original Dataset is split into Train and Test. The train dataset is used for training the Model while the test dataset is used to evaluate the model.

**Feature
Scaling**

Scaling helps in Interpretation. It is important to have all the categorical variables on the same scale to have better readability. Standardization brings the scaled data into standard normal distribution with mean equal to 0 and Standard Deviation is equal to 1.

FEATURE SELECTION USING RFE

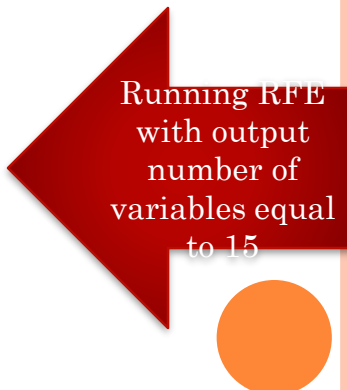
Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model's `coef_` or `feature_importances_` attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, 15)          # running RFE with 15 variables as output
rfe = rfe.fit(X_train, y_train)

col = X_train.columns[rfe.support_]
col

Index(['Do Not Email', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website',
       'What is your current occupation_Working Professional', 'Tags_Busy',
       'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_invalid number',
       'Tags_switched off', 'Tags_wrong number given', 'Lead Quality_Not Sure',
       'Lead Quality_Worst', 'Last Notable Activity_SMS Sent'],
      dtype='object')
```



Running RFE
with output
number of
variables equal
to 15

PREDICTING THE CONVERSION PROBABILITY AND PREDICTED COLUMN

*Creating dataframe with actual
Converted flag and Predicted
Probabilities*

*Fig shows top 5 records on the data
frame towards right*

	Converted	Converted_prob	LeadID
0	1	0.861689	160
1	0	0.185663	2267
2	1	0.990027	8895
3	1	0.861689	854
4	0	0.185663	3640

	Converted	Converted_prob	LeadID	predicted
0	1	0.861689	160	1
1	0	0.185663	2267	0
2	1	0.990027	8895	1
3	1	0.861689	854	1
4	0	0.185663	3640	0

*Creating new column 'predicted'
with 1 if Churn_Prob > 0.5 else 0*

*Fig shows top 5 records on the data
frame towards right*



OPTIMAL PROBABILITY THRESHOLD

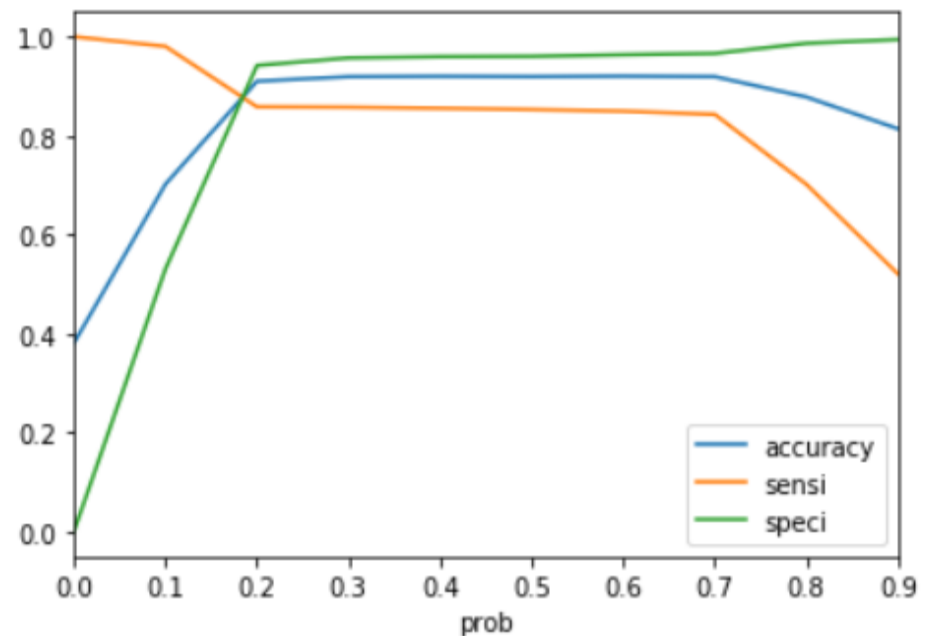
- Optimal cutoff probability is that prob where get balanced sensitivity and specificity.

Optimal Probability Threshold

➤ The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.

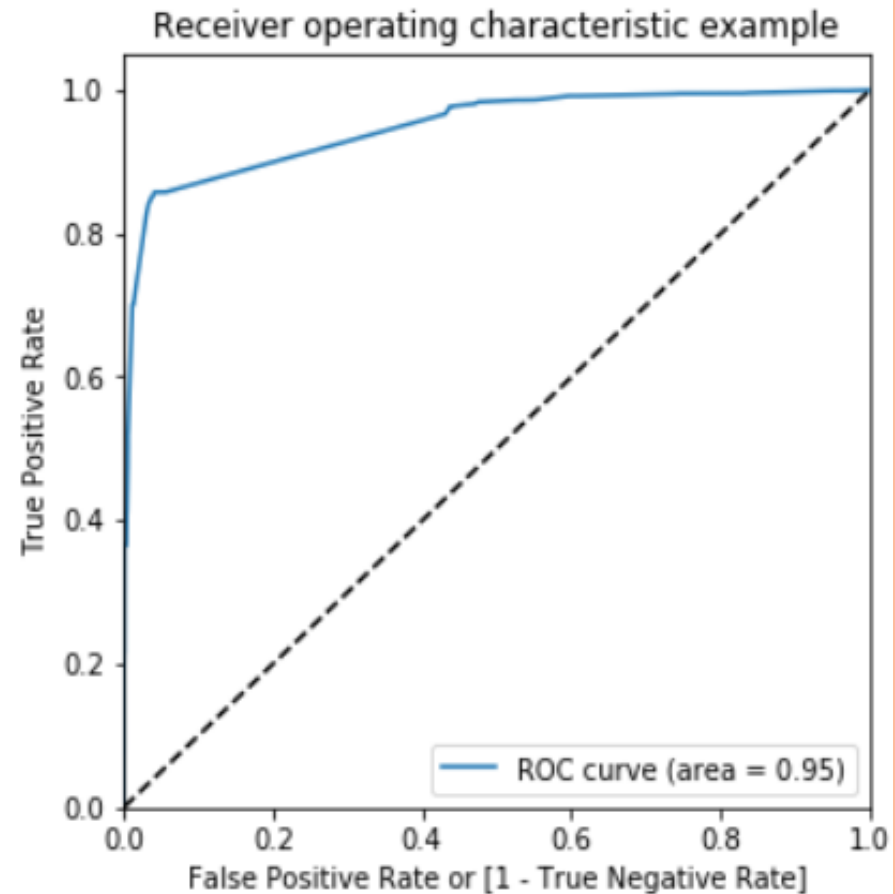
➤ From the curve above the optimum point for cutoff probability is found to be 0.2.

➤ At this point all the 3 metrics, accuracy sensitivity specificity was found to be more than 80% which is a well acceptable value.



ROC CURVE

- It shows the trade off between sensitivity and specificity
- Any increase in sensitivity will be accompanied by decrease in specificity.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- From the fig on the right the model curve is close to left-hand border indicating accuracy is more.



AREA UNDER CURVE AUC(GINI)

➤ By determining AUC of the ROC curve, the goodness of the model is determined.

➤ Since the ROC curve is more towards the upper-left corner of the graph it means that the model is very good.

➤ The value of AUC is 0.95 for our model which tells our model is excellent.

follows,

is to be doing well on the test dataset.



EVALUATING MODEL ON TRAIN DATASET

Confusion Matrix

#Predicted	not_converted	
converted	4316	179
#Actual	407	2357
not_converted		
converted		



**Probability
Threshold**
= 0.2

Accuracy
0.904

Sensitivity
0.843

Specificity
0.940

**False Positive
Rate**
0.039

**Positive
Predictive
Value**
0.929

**Negative
Predictive
Value**
0.9138

Precision
0.929

Recall
0.852

F1 Score
0.874

**Area
Under
Curve**
0.948



MAKING PREDICTIONS ON THE TEST SET

- The predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold of 0.2 the leads from the test were predicted if they will convert or not.

	LeadID	Converted	Converted_prob	final_predicted
0	3271	0	0.185663	0
1	1490	1	0.960771	1
2	7936	0	0.185663	0
3	4216	1	0.999083	1
4	3830	0	0.185663	0

Conversion Matrix

#Predicted	not_converted	converted
#Actual		
not_converted	1076	68
converted	105	566



EVALUATING MODEL ON TEST DATASET

Accuracy
0.904

Sensitivity
0.843

Specificity
0.940

False Positive
Rate
0.05

Positive
Predictive
Value
0.89

Negative
Predictive
Value
0.911

Precision
0.929

Recall
0.852

F1 Score
0.874

Area
Under
Curve
0.939

Cross
Validati
on
0.89

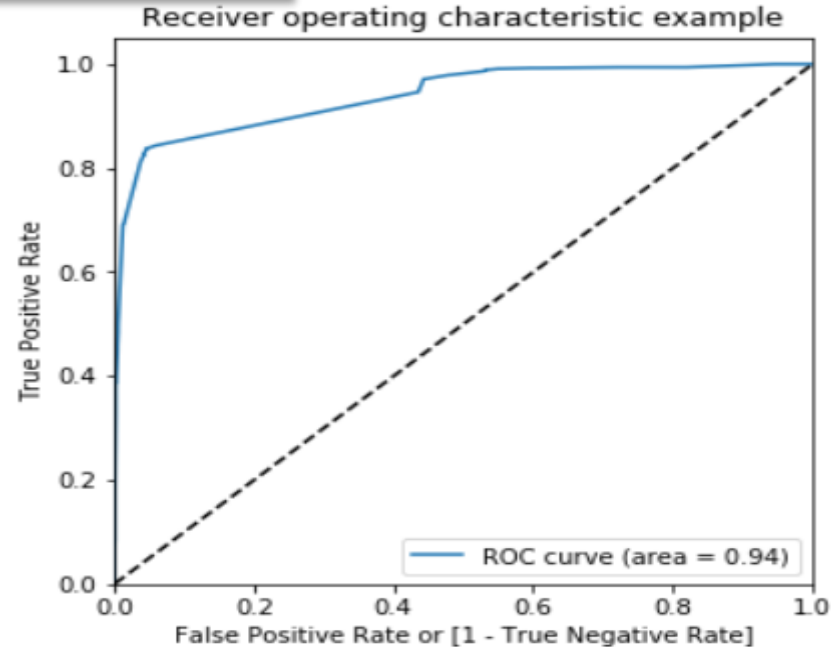


CLASSIFICATION REPORT AND ROC OF TEST DATA

Classification Report

	precision	recall	f1-score	support
0	0.91	0.94	0.93	1144
1	0.89	0.84	0.87	671

ROC Curve



LEAD SCORE CALCULATION

$$\text{Lead Score} = 100 * \text{Conversion Probability}$$

➤The test and train data's are concatenated to obtain the lead score.

➤Higher the lead score higher is the probability of conversion and vice versa.

➤Since we used 0.2 as threshold any lead with a lead score of 20 or above will have a value of 1 in final_predicted column.

	Lead Number	Converted	Converted_prob	final_predicted	Lead_Score
0	660737	0	0.112052	0	11
1	660728	0	0.000866	0	0
2	660727	1	0.861689	1	86
3	660719	0	0.000866	0	0
4	660681	1	0.861689	1	86
5	660680	0	0.185663	0	19
6	660673	1	0.861689	1	86
7	660664	0	0.185663	0	19
8	660624	0	0.185663	0	19
9	660616	0	0.185663	0	19

DETERMINING FEATURE IMPORTANCE

➤ *15 features have been used in our model to predict if a lead will get converted or not.*

➤ *The Beta Co-efficient values for these features from model parameters are used to determine the order of importance of these features.*

➤ *Features with high positive Beta are the ones that contribute most in determining the probability of lead conversion and low Beta values contribute the least.*

Do Not Email	-13.82
Lead Origin_Lead Add Form	11.73
Lead Source_Welingak Website	39.59
What is your current occupation_Working Professional	15.48
Tags_Busy	42.88
Tags_Closed by Horizzon	90.79
Tags_Lost to EINS	100.00
Tags_Ringing	-18.92
Tags_Will revert after reading the email	44.10
Tags_switched off	-27.83
Lead Quality_Not Sure	-37.41
Lead Quality_Worst	-44.08
Last Notable Activity_SMS Sent	31.31

DETERMINING FEATURE IMPORTANCE

➤ *The Relative Importance of each Feature is determined on a scale of 100, with the feature with highest importance having a scale of 100.*

➤ *$\text{Feature_Importance} = 100.0 * (\text{feature_importance} / \text{feature_importance.max()})$*

➤ *Features are then sorted using Quick Sort algorithm.*

➤ *Fig shows the sorted features plotted in bar graph in descending order of their relative Importance.*



INFERENCE

- After trying several model we finally choose the model with following characteristics

- *All variables have $p\text{-value} < 0.05$.*
- *All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features.*
- *This is also evident from the heat map.*
- *The overall accuracy of 0.904 at a probability threshold of 0.2 on the test dataset is also very acceptable*

Using the model, the dependent variable value was predicted as per the following threshold values of Conversion Probability:

Dataset	ThresholdValue	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive value	Precision	Recall	F1 Value	Cross Validation Score	AUC
Train	0.2	0.904	0.843	0.940	0.039	0.929	0.913	0.929	0.852	0.874		0.948
Test	0.2	0.90	0.834	0.940	0.05	0.89	0.911	0.929	0.852	0.876	0.89	0.939

INFERENCE

Features With Positive Coefficient Values:

- 1. Tags_Lost to EINS***
- 2. Tags_Closed by Horizzon***
- 3. Tags_Will revert after reading the email***
- 4. Lead Source_Welingak Website***
- 5. Last Acitivity_SMS Sent***
- 6. What is your current occupation_Working Professional***
- 7. What is your current occupation_Unemployed.***

The Conversion Probability of Lead decreases with decreases in in values of following features in descending Order

The Conversion Probability of Lead increases with increase in in values of following features in descending Order

Features with Negative Coefficient Values:

- 1. Tags_switched off***
- 2. Tags_Ringing***
- 3. Tags_Already a Student***
- 4. Tags_Not doing further education***
- 5. Lead Quality_worst***
- 6. Tags_opp hangup***
- 7. Tags_Interested in full time MBA***
- 8. Tags_Interested in other courses***
- 9. Assymetrique Activity Index_03 Low***

PROBLEM SOLUTION & RECOMMENDATIONS


The top 3 feature contributing most to lead conversion are:

- 1. Tags_Lost to EINS*
- 2. Tags_Closed by Horizzon*
- 3. Tags_Will revert after reading the email*

The top 3 categorical/dummy variables contributing most to lead conversion are:

- 1. Tags_Lost to EINS*
- 2. Tags_Closed by Horizzon*
- 3. Tags_Will revert after reading the email*

The Threshold values for Conversion probability ha to be chosen wisely. Which will in turn effect the sensitivity(lower threshold) and specificity(Higher threshold) there by ensuring that all the leads are converted.



Thank You

